EXPRESSED SEQUENCES OF *ARABIDOPSIS THALIANA*

CROSS –REFERENCE TO RELATED APPLICATION

This application claims the benefit of U.S. application serial no. 09/638,258 and U.S. Provisional Application 60/148,784 filed August 13, 1999.

FIELD OF INVENTION

The invention is in the field of polynucleotide sequences of a plant, particularly sequences expressed in arabidopsis thaliana.

BACKGROUND OF THE INVENTION

Plants and plant products have vast commercial importance in a wide variety of areas including food crops for human and animal consumption, flavor enhancers for food, and production of specialty chemicals for use in products such as medicaments and fragrances. In considering food crops for humans and livestock, genes such as those involved in a plant's resistance to insects, plant viruses, and fungi; genes involved in pollination; and genes whose products enhance the nutritional value of the food, are of major importance. A number of such genes have been described, see, for example, McCaskill and Croteau (1999) Nature Biotechnol. 17:31-36.

Despite recent advances in methods for identification, cloning, and characterization of genes, much remains to be learned about plant physiology in general, including how plants produce many of the above-mentioned products; mechanisms for resistance to herbicides, insects, plant viruses, fungi; elucidation of genes involved in specific biosynthetic pathways; and genes involved in environmental tolerance, *e.g.*, salt tolerance, drought tolerance, or tolerance to anaerobic conditions.

Arabidopsis thaliana is a model system for genetic, molecular and biochemical studies of higher plants. Features of this plant that make it a model system for genetic and molecular biology research include a small genome size, organized into five chromosomes and containing an estimated 20,000 genes, a rapid life cycle, prolific seed production and, since it is small, it can easily be cultivation in limited

space. *A. thaliana* is a member of the mustard family (*Brassicaceae*) with a broad natural distribution throughout Europe, Asia, and North America. Many different ecotypes have been collected from natural populations and are available for experimental analysis. The entire life cycle, including seed germination, formation of a rosette plant, bolting of the main stem, flowering, and maturation of the first seeds, is completed in 6 weeks. A large number of mutant lines are available that affect nearly all aspects of its growth. These features greatly facilitate the isolation of fundamentally interesting and potentially important genes for agronomic development

Most gene products from higher plants exhibit adequate sequence similarity to deduced amino acid sequences of other plant genes to permit assignment of probable gene function, if it is known, in any higher plant. It is likely that there will be very few protein-encoding angiosperm genes that do not have orthologs or paralogs in *Arabidopsis*. The developmental diversity of higher plants may be largely due to changes in the cis-regulatory sequences of transcriptional regulators and not in coding sequences.

Many advances reported over the past few years offer clear evidence that this plant is not only a very important model species for basic research, but also extremely valuable for applied plant scientists and plant breeders. Knowledge gained from *Arabidopsis* can be used directly to develop desired traits in plants of other species.

Relevant Literature

Cold Spring Harbor Monograph 27 (1994) E.M. Meyerowitz and C.R. Somerville, eds. (CSH Laboratory Press). Annual Plant Reviews, Vol. 1: Arabidopsis (1998) M. Anderson and J.A. Roberts, eds. (CRC Press). Methods in Molecular Biology: Arabidopsis Protocols, Vol. 82 (1997) J.M. Martinez-Zapater and J. Salinas, eds. (CRC Press).

Mayer *et al* (1999) Nature **402**(6763):769-77; "Sequence and analysis of chromosome 4 of the plant Arabidopsis thaliana". Lin *et al.* (1999) **402**(6763):761-8, "Sequence and analysis of chromosome 2 of the plant Arabidopsis thaliana". Meinke *et al.* (1998) Science **282**:662-682, "*Arabidopsis thaliana*: a model plant for genome analysis". Somerville and Somerville (1999) Science **285**:380-383, "Plant functional

genomics". Mozo *et al.* (1999) Nat. Genet. **22**:271-275, "A complete BAC-based physical map of the *Arabidopsis thaliana* genome".

SUMMARY OF THE INVENTION

Novel nucleic acid sequences of *Arabidopsis thaliana,* their encoded polypeptides and variants thereof, genes corresponding to these nucleic acids, and proteins expressed by the genes, are provided.

The invention also provides diagnostic, prophylactic and therapeutic agents employing such novel nucleic acids, their corresponding genes or gene products, including expression constructs, probes, antisense constructs, and the like. The genetic sequences may also be used for the genetic manipulation of plant cells, particularly dicotyledonous plants. The encoded gene products and modified organisms are useful for introducing or improving disease resistance and stress tolerance into plants; screening of biologically active agents, *e.g.* fungicides, *etc.*; for elucidating biochemical pathways; and the like.

In one embodiment of the invention, a nucleic acid is provided that comprises a start codon; an optional intervening sequence; a coding sequence capable of hybridizing under stringent conditions as set forth in SEQ ID NO:1 to 900; and an optional terminal sequence, wherein at least one of said optional sequences is present. Such a nucleic acid may correspond to naturally occurring *Arabidopsis* expressed sequences.

DETAILED DESCRIPTION OF THE INVENTION

Novel nucleic acid sequences from *Arabidopsis thaliana*, their encoded polypeptides and variants thereof, genes corresponding to these nucleic acids and proteins expressed by the genes are provided. The invention also provides agents employing such novel nucleic acids, their corresponding genes or gene products, including expression constructs, probes, antisense constructs, and the like. The nucleotide sequences are provided in the attached SEQLIST.

Sequences include, but are not limited to, sequences that encode resistance proteins; sequences that encode tolerance factors; sequences encoding proteins or other factors that are involved, directly or indirectly in biochemical pathways such as metabolic or biosynthetic pathways, sequences involved in signal transduction,

sequences involved in the regulation of gene expression, structural genes, and the like. Biosynthetic pathways of interest include, but are not limited to, biosynthetic pathways whose product (which may be an end product or an intermediate) is of commercial, nutritional, or medicinal value.

The sequences may be used in screening assays of various plant strains to determine the strains that are best capable of withstanding a particular disease or environmental stress. Sequences encoding activators and resistance proteins may be introduced into plants that are deficient in these sequences. Alternatively, the sequences may be introduced under the control of promoters that are convenient for induction of expression. The protein products may be used in screening programs for insecticides, fungicides and antibiotics to determine agents that mimic or enhance the resistance proteins. Such agents may be used in improved methods of treating crops to prevent or treat disease. The protein products may also be used in screening programs to identify agents which mimic or enhance the action of tolerance factors. Such agents may be used in improved methods of treating crops to enhance their tolerance to environmental stresses.

Still other embodiments of the invention provide methods for enhancing or inhibiting production of a biosynthetic product in a plant by introducing a nucleic acid of the invention into a plant cell, where the nucleic acid comprises sequences encoding a factor which is involved, directly or indirectly in a biosynthetic pathway whose products are of commercial, nutritional, or medicinal value include any factor, usually a protein or peptide, which regulates such a biosynthetic pathway; which is an intermediate in such a biosynthetic pathway; or which in itself is a product that increases the nutritional value of a food product; or which is a medicinal product; or which is any product of commercial value.

Transgenic plants containing the antisense nucleic acids of the invention are useful for identifying other mediators that may induce expression of proteins of interest; for establishing the extent to which any specific insect and/or pathogen is responsible for damage of a particular plant; for identifying other mediators that may enhance or induce tolerance to environmental stress; for identifying factors involved in biosynthetic pathways of nutritional, commercial, or medicinal value; or for identifying products of nutritional, commercial, or medicinal value.

In still other embodiments, the invention provides transgenic plants constructed by introducing a subject nucleic acid of the invention into a plant cell, and growing the cell into a callus and then into a plant; or, alternatively by breeding a transgenic plant from the subject process with a second plant to form an F1 or higher hybrid. The subject transgenic plants and progeny are used as crops for their enhanced disease resistance, enhanced traits of interest, for example size or flavor of fruit, length of growth cycle, *etc.*, or for screening programs, *e.g.* to determine more effective insecticides, *etc*; used as crops which exhibit enhanced tolerance environmental stress; or used to produce a factor.

Those skilled in the art will recognize the agricultural advantages inherent in plants constructed to have either increased or decreased expression of resistance proteins; or increased or decreased tolerance to environmental factors; or which produce or over-produce one or more factors involved in a biosynthetic pathway whose product is of commercial, nutritional, or medicinal value. For example, such plants may have increased resistance to attack by predators, insects, pathogens, microorganisms, herbivores, mechanical damage and the like; may be more tolerant to environmental stress, *e.g.* may be better able to withstand drought conditions, freezing, and the like; or may produce a product not normally made in the plant, or may produce a product in higher than normal amounts, where the product has commercial, nutritional, or medicinal value. Plants which may be useful include dicotyledons and monocotyledons. Representative examples of plants in which the provided sequences may be useful include tomato, potato, tobacco, cotton, soybean, alfalfa, rape, and the like. Monocotyledons, more particularly grasses (*Poaceae* family) of interest, include, without limitation, *Avena sativa* (oat); *Avena strigosa* (black oat); *Elymus* (wild rye); *Hordeum sp.* including *Hordeum vulgare* (barley); *Oryza sp.*, including *Oryza glaberrima* (African rice); *Oryza longistaminata* (long-staminate rice); *Pennisetum americanum* (pearl millet); *Sorghum sp.* (sorghum); *Triticum sp.*, including *Triticum aestivum* (common wheat); *Triticum durum* (durum wheat); *Zea mays* (corn); *etc.*

NUCLEIC ACID COMPOSITIONS

The following detailed description describes the nucleic acid compositions encompassed by the invention, methods for obtaining cDNA or genomic DNA

encoding a full-length gene product, expression of these nucleic acids and genes; identification of structural motifs of the nucleic acids and genes; identification of the function of a gene product encoded by a gene corresponding to a nucleic acid of the invention; use of the provided nucleic acids as probes, in mapping, and in diagnosis; use of the corresponding polypeptides and other gene products to raise antibodies; use of the nucleic acids in genetic modification of plant and other species; and use of the nucleic acids, their encoded gene products, and modified organisms, for screening and diagnostic purposes.

The scope of the invention with respect to nucleic acid compositions includes, but is not necessarily limited to, nucleic acids having a sequence set forth in any one of SEQ ID NOS:1-900; nucleic acids that hybridize the provided sequences under stringent conditions; genes corresponding to the provided nucleic acids; variants of the provided nucleic acids and their corresponding genes, particularly those variants that retain a biological activity of the encoded gene product.

In one embodiment, the sequences of the invention provide a polypeptide coding sequence. The polypeptide coding sequence may correspond to a naturally expressed mRNA in Arabidopsis or other species, or may encode a fusion protein between one of the provided sequences and an exogenous protein coding sequence. The coding sequence is characterized by an ATG start codon, a lack of stop codons in-frame with the ATG, and a termination codon, that is, a continuous open frame is provided between the start and the stop codon. The sequence contained between the start and the stop codon will comprise a sequence capable of hybridizing under stringent conditions to a sequence set for in SEQ ID NO:1-900, and may comprise the sequence set forth in the Seqlist.

Other nucleic acid compositions contemplated by and within the scope of the present invention will be readily apparent to one of ordinary skill in the art when provided with the disclosure here.

The invention features nucleic acids that are derived from *Arabidopsis thaliana*. Novel nucleic acid compositions of the invention of particular interest comprise a sequence set forth in any one of SEQ ID NOS:1-900 or an identifying sequence thereof. An "identifying sequence" is a contiguous sequence of residues at least about 10 nt to about 20 nt in length, usually at least about 50 nt to about 100 nt in length, that uniquely identifies a nucleic acid sequence, e.g., exhibits less

than 90%, usually less than about 80% to about 85% sequence identity to any contiguous nucleotide sequence of more than about 20 nt. Thus, the subject novel nucleic acid compositions include full length cDNAs or mRNAs that encompass an identifying sequence of contiguous nucleotides from any one of SEQ ID NOS:1-900.

The nucleic acids of the invention also include nucleic acids having sequence similarity or sequence identity. Nucleic acids having sequence similarity are detected by hybridization under low stringency conditions, for example, at 50°C and 10XSSC (0.9 M NaCl/0.09 M sodium citrate) and remain bound when subjected to washing at 55°C in 1XSSC. Sequence identity can be determined by hybridization under stringent conditions, for example, at 50°C or higher and 0.1XSSC (9 mM NaCl/0.9 mM sodium citrate). Hybridization methods and conditions are well known in the art, see U.S. Patent No. 5,707,829. Nucleic acids that are substantially identical to the provided nucleic acid sequences, *e.g.* allelic variants, genetically altered versions of the gene, *etc.*, bind to the provided nucleic acid sequences (SEQ ID NOS:1-900) under stringent hybridization conditions. By using probes, particularly labeled probes of DNA sequences, one can isolate homologous or related genes. The source of homologous genes can be any species, particularly grasses as previously described.

Preferably, hybridization is performed using at least 15 contiguous nucleotides of at least one of SEQ ID NOS:1-900. The probe will preferentially hybridize with a nucleic acid or mRNA comprising the complementary sequence, allowing the identification and retrieval of the nucleic acids of the biological material that uniquely hybridize to the selected probe. Probes of more than 15 nucleotides can be used, e.g. probes of from about 18 nucleotides up to the entire length of the provided nucleic acid sequences, but 15 nucleotides generally represents sufficient sequence for unique identification.

The nucleic acids of the invention also include naturally occurring variants of the nucleotide sequences, *e.g.* degenerate variants, allelic variants, *etc.* Variants of the nucleic acids of the invention are identified by hybridization of putative variants with nucleotide sequences disclosed herein, preferably by hybridization under stringent conditions For example, by using appropriate wash conditions, variants of the nucleic acids of the invention can be identified where the allelic variant exhibits at most about 25-30% base pair mismatches relative to the selected nucleic acid

probe. In general, allelic variants contain 5-25% base pair mismatches, and can contain as little as even 2-5%, or 1-2% base pair mismatches, as well as a single base-pair mismatch.

The invention also encompasses homologs corresponding to the nucleic acids of SEQ ID NOS:1-900, where the source of homologous genes can be any related species, usually within the same genus or group. Homologs have substantial sequence similarity, *e.g.* at least 75% sequence identity, usually at least 90%, more usually at least 95% between nucleotide sequences. Sequence similarity is calculated based on a reference sequence, which may be a subset of a larger sequence, such as a conserved motif, coding region, flanking region, etc. A reference sequence will usually be at least about 18 contiguous nt long, more usually at least about 30 nt long, and may extend to the complete sequence that is being compared. Algorithms for sequence analysis are known in the art, such as BLAST, described in Altschul et al., J. Mol. Biol. (1990) 215:403-10.

In general, variants of the invention have a sequence identity greater than at least about 65%, preferably at least about 75%, more preferably at least about 85%, and can be greater than at least about 90% or more as determined by the Smith-Waterman homology search algorithm as implemented in MPSRCH program (Oxford Molecular). For the purposes of this invention, a preferred method of calculating percent identity is the Smith-Waterman algorithm, using the following. Global DNA sequence identity must be greater than 65% as determined by the Smith-Wateman homology search algorithm as implemented in MPSRCH program (Oxford Molecular) using an affine gap search with the following search parameters: gap open penalty, 12; and gap extention penalty, 1.

The subject nucleic acids can be cDNAs or genomic DNAs, as well as fragments thereof, particularly fragments that encode a biologically active gene product and/or are useful in the methods disclosed herein. The term "cDNA" as used herein is intended to include all nucleic acids that share the arrangement of sequence elements found in native mature mRNA species, where sequence elements are exons and 3' and 5' non-coding regions. Normally mRNA species have contiguous exons, with the introns, when present, being removed by nuclear RNA splicing, to create a continuous open reading frame encoding a polypeptide of the invention.

A genomic sequence of interest comprises the nucleic acid present between the initiation codon and the stop codon, as defined in the listed sequences, including all of the introns that are normally present in a native chromosome. It can further include the 3' and 5' untranslated regions found in the mature mRNA. It can further include specific transcriptional and translational regulatory sequences, such as promoters, enhancers, etc., including about 1 kb, but possibly more, of flanking genomic DNA at either the 5' and 3' end of the transcribed region. The genomic DNA can be isolated as a fragment of 100 kb or smaller; and substantially free of flanking chromosomal sequence. The genomic DNA flanking the coding region, either 3' and 5', or internal regulatory sequences as sometimes found in introns, contains sequences required for expression.

The nucleic acid compositions of the subject invention can encode all or a part of the subject expressed polypeptides. Double or single stranded fragments can be obtained from the DNA sequence by chemically synthesizing oligonucleotides in accordance with conventional methods, by restriction enzyme digestion, by PCR amplification, etc. Isolated nucleic acids and nucleic acid fragments of the invention comprise at least about 15 up to about 100 contiguous nucleotides, or up to the complete sequence provided in SEQ ID NOS:1-900. For the most part, fragments will be of at least 15 nt, usually at least 18 nt or 25 nt, and up to at least about 50 contiguous nt in length or more.

Probes specific to the nucleic acids of the invention can be generated using the nucleic acid sequences disclosed in SEQ ID NOS:1-900 and the fragments as described above. The probes can be synthesized chemically or can be generated from longer nucleic acids using restriction enzymes. The probes can be labeled, for example, with a radioactive, biotinylated, or fluorescent tag. Preferably, probes are designed based upon an identifying sequence of a nucleic acid of one of SEQ ID NOS:1-900. More preferably, probes are designed based on a contiguous sequence of one of the subject nucleic acids that remain unmasked following application of a masking program for masking low complexity (e.g., XBLAST) to the sequence., *i.e.* one would select an unmasked region, as indicated by the nucleic acids outside the poly-n stretches of the masked sequence produced by the masking program.

The nucleic acids of the subject invention are isolated and obtained in substantial purity, generally as other than an intact chromosome. Usually, the

nucleic acids, either as DNA or RNA, will be obtained substantially free of other naturally-occurring nucleic acid sequences, generally being at least about 50%, usually at least about 90% pure and are typically "recombinant", e.g., flanked by one or more nucleotides with which it is not normally associated on a naturally occurring chromosome.

The nucleic acids of the invention can be provided as a linear molecule or within a circular molecule. They can be provided within autonomously replicating molecules (vectors) or within molecules without replication sequences. They can be regulated by their own or by other regulatory sequences, as is known in the art. The nucleic acids of the invention can be introduced into suitable host cells using a variety of techniques which are available in the art, such as transferrin polycation-mediated DNA transfer, transfection with naked or encapsulated nucleic acids, liposome-mediated DNA transfer, intracellular transportation of DNA-coated latex beads, protoplast fusion, viral infection, electroporation, gene gun, calcium phosphate-mediated transfection, and the like.

The subject nucleic acid compositions can be used to, for example, produce polypeptides, as probes for the detection of mRNA of the invention in biological samples, *e.g.* extracts of cells, to generate additional copies of the nucleic acids, to generate ribozymes or antisense oligonucleotides, and as single stranded DNA probes or as triple-strand forming oligonucleotides. The probes described herein can be used to, for example, determine the presence or absence of the nucleic acid sequences as shown in SEQ ID NOS:1-900 or variants thereof in a sample. These and other uses are described in more detail below.

USE OF NUCLEIC ACIDS AS CODING SEQUENCES

Naturally occurring Arabidopsis polypeptides or fragments thereof are encoded by the provided nucleic acids. Methods are known in the art to determine whether the complete native protein is encoded by a candidate nucleic acid sequence. Where the provided sequence encodes a fragment of a polypeptide, methods known in the art may be used to determine the remaining sequence. These approaches may utilize a bioinformatics approach, a cloning approach, extension of mRNA species, *etc.*

Substantial genomic sequence is available for Arabidopsis, and may be exploited for determining the complete coding sequence corresponding to the provided sequences. The region of the chromosome to which a given sequence is located may be determined by hybridization or by database searching. The genomic sequence is then searched upstream and downstream for the presence of intron/exon boundaries, and for motifs characteristic of transcriptional start and stop sequences, for example by using Genscan (Burge and Karlin (1997) J. Mol. Biol. **268**:78-94); or GRAIL (Uberbacher and Mural (1991) P.N.A.S. **88**:11261-1265).

Alternatively, nucleic acid having a sequence of one of SEQ ID NOS:1-900, or an identifying fragment thereof, is used as a hybridization probe to complementary molecules in a cDNA library using probe design methods, cloning methods, and clone selection techniques as known in the art. Libraries of cDNA are made from selected cells. The cells may be those of *A. thaliana*, or of related species. In some cases it will be desirable to select cells from a particular stage, *e.g.* seeds, leaves, infected cells, *etc.*

Techniques for producing and probing nucleic acid sequence libraries are described, for example, in Sambrook et al., Molecular Cloning: A Laboratory Manual, 2nd Ed., (1989) Cold Spring Harbor Press, Cold Spring Harbor, NY; and Current Protocols in Molecular Biology, (1987 and updates) Ausubel et al., eds. The cDNA can be prepared by using primers based on sequence from SEQ ID NOS:1-900. In one embodiment, the cDNA library can be made from only poly-adenylated mRNA. Thus, poly-T primers can be used to prepare cDNA from the mRNA.

Members of the library that are larger than the provided nucleic acids, and preferably that encompass the complete coding sequence of the native message, are obtained. In order to confirm that the entire cDNA has been obtained, RNA protection experiments are performed as follows. Hybridization of a full-length cDNA to an mRNA will protect the RNA from RNase degradation. If the cDNA is not full length, then the portions of the mRNA that are not hybridized will be subject to RNase degradation. This is assayed, as is known in the art, by changes in electrophoretic mobility on polyacrylamide gels, or by detection of released monoribonucleotides. Sambrook et al., Molecular Cloning: A Laboratory Manual, 2nd Ed., (1989) Cold Spring Harbor Press, Cold Spring Harbor, NY. In order to obtain

additional sequences 5' to the end of a partial cDNA, 5' RACE (PCR Protocols: A Guide to Methods and Applications, (1990) Academic Press, Inc.) may be performed.

Genomic DNA is isolated using the provided nucleic acids in a manner similar to the isolation of full-length cDNAs. Briefly, the provided nucleic acids, or portions thereof, are used as probes to libraries of genomic DNA. Preferably, the library is obtained from the cell type that was used to generate the nucleic acids of the invention, but this is not essential. Such libraries can be in vectors suitable for carrying large segments of a genome, such as P1 or YAC, as described in detail in Sambrook et al., 9.4-9.30. In order to obtain additional 5' or 3' sequences, chromosome walking is performed, as described in Sambrook *et al.*, such that adjacent and overlapping fragments of genomic DNA are isolated. These are mapped and pieced together, as is known in the art, using restriction digestion enzymes and DNA ligase.

PCR methods may be used to amplify the members of a cDNA library that comprise the desired insert. In this case, the desired insert will contain sequence from the full length cDNA that corresponds to the instant nucleic acids. Such PCR methods include gene trapping and RACE methods. Gene trapping entails inserting a member of a cDNA library into a vector. The vector then is denatured to produce single stranded molecules. Next, a substrate-bound probe, such a biotinylated oligo, is used to trap cDNA inserts of interest. Biotinylated probes can be linked to an avidin-bound solid substrate. PCR methods can be used to amplify the trapped cDNA. To trap sequences corresponding to the full length genes, the labeled probe sequence is based on the nucleic acid sequences of the invention. Random primers or primers specific to the library vector can be used to amplify the trapped cDNA. Such gene trapping techniques are described in Gruber *et al.*, WO 95/04745 and Gruber *et al.*, U.S. Pat. No. 5,500,356. Kits are commercially available to perform gene trapping experiments from, for example, Life Technologies, Gaithersburg, Maryland, USA.

"Rapid amplification of cDNA ends", or RACE, is a PCR method of amplifying cDNAs from a number of different RNAs. The cDNAs are ligated to an oligonucleotide linker, and amplified by PCR using two primers. One primer is based on sequence from the instant nucleic acids, for which full length sequence is desired, and a second primer comprises sequence that hybridizes to the oligonucleotide linker

to amplify the cDNA. A description of this methods is reported in WO 97/19110. A common primer may be designed to anneal to an arbitrary adaptor sequence ligated to cDNA ends. When a single gene-specific RACE primer is paired with the common primer, preferential amplification of sequences between the single gene specific primer and the common primer occurs. Commercial cDNA pools modified for use in RACE are available.

Once the full-length cDNA or gene is obtained, DNA encoding variants can be prepared by site-directed mutagenesis, described in detail in Sambrook et al., 15.3-15.63. The choice of codon or nucleotide to be replaced can be based on disclosure herein on optional changes in amino acids to achieve altered protein structure and/or function. As an alternative method to obtaining DNA or RNA from a biological material, nucleic acid comprising nucleotides having the sequence of one or more nucleic acids of the invention can be synthesized.

EXPRESSION OF POLYPEPTIDES

The provided nucleic acid, *e.g.* a nucleic acid having a sequence of one of SEQ ID NOS:1-900), the corresponding cDNA, the polypeptide coding sequence as described above, or the full-length gene is used to express a partial or complete gene product. Constructs of nucleic acids having sequences of SEQ ID NOS:1-900 can be generated by recombinant methods, synthetically, or in a single-step assembly of a gene and entire plasmid from large numbers of oligodeoxyribonucleotides is described by, *e.g.* Stemmer *et al.*, Gene (Amsterdam) (1995) 164(1):49-53.

Appropriate nucleic acid constructs are purified using standard recombinant DNA techniques as described in, for example, Sambrook et al., Molecular Cloning: A Laboratory Manual, 2nd Ed., (1989) Cold Spring Harbor Press, Cold Spring Harbor, NY. The gene product encoded by a nucleic acid of the invention is expressed in any expression system, including, for example, bacterial, yeast, insect, amphibian and mammalian systems.

The subject nucleic acid molecules are generally propagated by placing the molecule in a vector. Viral and non-viral vectors are used, including plasmids. The choice of plasmid will depend on the type of cell in which propagation is desired and the purpose of propagation. Certain vectors are useful for amplifying and making

large amounts of the desired DNA sequence. Other vectors are suitable for expression in cells in culture. Still other vectors are suitable for transfer and expression in cells in a whole organism or person. The choice of appropriate vector is well within the skill of the art. Many such vectors are available commercially.

The nucleic acids set forth in SEQ ID NOS:1-900 or their corresponding full-length nucleic acids are linked to regulatory sequences as appropriate to obtain the desired expression properties. These can include promoters attached either at the 5' end of the sense strand or at the 3' end of the antisense strand, enhancers, terminators, operators, repressors, and inducers. The promoters can be regulated or constitutive. In some situations it may be desirable to use conditionally active promoters, such as tissue-specific or developmental stage-specific promoters. These are linked to the desired nucleotide sequence using the techniques described above for linkage to vectors. Any techniques known in the art can be used.

When any of the above host cells, or other appropriate host cells or organisms, are used to replicate and/or express the nucleic acids or nucleic acids of the invention, the resulting replicated nucleic acid, RNA, expressed protein or polypeptide, is within the scope of the invention as a product of the host cell or organism. The product is recovered by any appropriate means known in the art.

IDENTIFICATION OF FUNCTIONAL AND STRUCTURAL MOTIFS

Translations of the nucleotide sequence of the provided nucleic acids, cDNAs or full genes can be aligned with individual known sequences. Similarity with individual sequences can be used to determine the activity of the polypeptides encoded by the nucleic acids of the invention. Also, sequences exhibiting similarity with more than one individual sequence can exhibit activities that are characteristic of either or both individual sequences.

The six possible reading frames may be translated using programs such as GCG pepdata, or GCG Frames (Wisconsin Package Version 10.0, Genetics Computer Group (GCG) , Madison, Wisconsin, USA.). Programs such as ORFFinder (National Center for Biotechnology Information (NCBI) a division of the National Library of Medicine (NLM) at the National Institutes of Health (NIH) http://www.ncbi.nlm.nih.gov/) may be used to identify open reading frames (ORFs) in sequences. ORF finder identifies all possible ORFs in a DNA sequence by

locating the standard and alternative stop and start codons. Other ORF identification programs include Genie (Kulp *et al.* (1996).

A generalized Hidden Markov Model may be used for the recognition of genes in DNA. (ISMB-96, St. Louis, MO, AAAI/MIT Press; Reese *et al.* (1997), "Improved splice site detection in Genie". Proceedings of the First Annual International Conference on Computational Molecular Biology RECOMB 1997, Santa Fe, NM, ACM Press, New York., P. 34.); BESTORF --Prediction of potential coding fragment in human or plant EST/mRNA sequence data using Markov Chain Models; and FGENEP -- Multiple genes structure prediction in plant genomic DNA (Solovyev *et al.* (1995) Identification of human gene structure using linear discriminant functions and dynamic programming. In Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology eds. Rawling *et al.* Cambridge, England, AAAI Press,367-375.; Solovyev *et al.* (1994) Nucl. Acids Res. **22**(24):5156-5163; Solovyev *et al,*. The prediction of human exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames, in: The Second International conference on Intelligent systems for Molecular Biology (eds. Altman et al.), AAAI Press, Menlo Park, CA (1994, 354-362) Solovyev and Lawrence, Prediction of human gene structure using dynamic programming and oligonucleotide composition, In: Abstracts of the 4th annual Keck symposium. Pittsburgh, 47,1993; Burge and Karlin (1997) J. Mol. Biol. **268**:78-94; Kulp *et al.* (1996) Proc. Conf. on Intelligent Systems in Molecular Biology '96, 134-142).

The full length sequences and fragments of the nucleic acid sequences of the nearest neighbors can be used as probes and primers to identify and isolate the full length sequence corresponding to provided nucleic acids. Typically, a selected nucleic acid is translated in all six frames to determine the best alignment with the individual sequences. These amino acid sequences are referred to, generally, as query sequences, which are aligned with the individual sequences. Suitable databases include Genbank, EMBL, and DNA Database of Japan (DDBJ).

Query and individual sequences can be aligned using the methods and computer programs described above, and include BLAST, available by ftp at ftp://ncbi.nlm.nih.gov/.

Gapped BLAST and PSI-BLAST are useful search tools provided by NCBI. (version 2.0) (Altschul *et al.*, 1997). Position-Specific Iterated BLAST (PSI-BLAST)

provides an automated, easy-to-use version of a "profile" search, which is a sensitive way to look for sequence homologues. The program first performs a gapped BLAST database search. The PSI-BLAST program uses the information from any significant alignments returned to construct a position-specific score matrix, which replaces the query sequence for the next round of database searching. PSI-BLAST may be iterated until no new significant alignments are found. The Gapped BLAST algorithm allows gaps (deletions and insertions) to be introduced into the alignments that are returned. Allowing gaps means that similar regions are not broken into several segments. The scoring of these gapped alignments tends to reflect biological relationships more closely. The Smith-Waterman is another algorithm that produces local or global gapped sequence alignments, see Meth. Mol. Biol. (1997) 70: 173-187. Also, the GAP program using the Needleman and Wunsch global alignment method can be utilized for sequence alignments.

Results of individual and query sequence alignments can be divided into three categories, high similarity, weak similarity, and no similarity. Individual alignment results ranging from high similarity to weak similarity provide a basis for determining polypeptide activity and/or structure. Parameters for categorizing individual results include: percentage of the alignment region length where the strongest alignment is found, percent sequence identity, and e value.

The percentage of the alignment region length is calculated by counting the number of residues of the individual sequence found in the region of strongest alignment, *e.g.* contiguous region of the individual sequence that contains the greatest number of residues that are identical to the residues of the corresponding region of the aligned query sequence. This number is divided by the total residue length of the query sequence to calculate a percentage. For example, a query sequence of 20 amino acid residues might be aligned with a 20 amino acid region of an individual sequence. The individual sequence might be identical to amino acid residues 5, 9-15, and 17-19 of the query sequence. The region of strongest alignment is thus the region stretching from residue 9-19, an 11 amino acid stretch. The percentage of the alignment region length is: 11 (length of the region of strongest alignment) divided by (query sequence length) 20 or 55%.

Percent sequence identity is calculated by counting the number of amino acid matches between the query and individual sequence and dividing total number of

matches by the number of residues of the individual sequences found in the region of strongest alignment. Thus, the percent identity in the example above would be 10 matches divided by 11 amino acids, or approximately, 90.9%

E value is the probability that the alignment was produced by chance. For a single alignment, the e value can be calculated according to Karlin et al., Proc. Natl. Acad. Sci. (1990) 87:2264 and Karlin et al., Proc. Natl. Acad. Sci. (1993) 90. The e value of multiple alignments using the same query sequence can be calculated using an heuristic approach described in Altschul et al., Nat. Genet. (1994) 6:119. Alignment programs such as BLAST program can calculate the e value.

Another factor to consider for determining identity or similarity is the location of the similarity or identity. Strong local alignment can indicate similarity even if the length of alignment is short. Sequence identity scattered throughout the length of the query sequence also can indicate a similarity between the query and profile sequences. The boundaries of the region where the sequences align can be determined according to Doolittle, *supra*; BLAST or FASTA programs; or by determining the area where sequence identity is highest.

In general, in alignment results considered to be of high similarity, the percent of the alignment region length is typically at least about 55% of total length query sequence; more typically, at least about 58%; even more typically; at least about 60% of the total residue length of the query sequence. Usually, percent length of the alignment region can be as much as about 62%; more usually, as much as about 64%; even more usually, as much as about 66%. Further, for high similarity, the region of alignment, typically, exhibits at least about 75% of sequence identity; more typically, at least about 78%; even more typically; at least about 80% sequence identity. Usually, percent sequence identity can be as much as about 82%; more usually, as much as about 84%; even more usually, as much as about 86%.

The p value is used in conjunction with these methods. The query sequence is considered to have a high similarity with a profile sequence when the p value is less than or equal to 10^{-2}. Confidence in the degree of similarity between the query sequence and the profile sequence increases as the p value become smaller.

In general, where alignment results considered to be of weak similarity, there is no minimum percent length of the alignment region nor minimum length of alignment. A better showing of weak similarity is considered when the region of

alignment is, typically, at least about 15 amino acid residues in length; more typically, at least about 20; even more typically; at least about 25 amino acid residues in length. Usually, length of the alignment region can be as much as about 30 amino acid residues; more usually, as much as about 40; even more usually, as much as about 60 amino acid residues. Further, for weak similarity, the region of alignment, typically, exhibits at least about 35% of sequence identity; more typically, at least about 40%; even more typically; at least about 45% sequence identity. Usually, percent sequence identity can be as much as about 50%; more usually, as much as about 55%; even more usually, as much as about 60%.

The query sequence is considered to have a low similarity with a profile sequence when the p value is greater than 10^{-2}. Confidence in the degree of similarity between the query sequence and the profile sequence decreases as the p values become larger.

Sequence identity alone can be used to determine similarity of a query sequence to an individual sequence and can indicate the activity of the sequence. Such an alignment, preferably, permits gaps to align sequences. Typically, the query sequence is related to the profile sequence if the sequence identity over the entire query sequence is at least about 15%; more typically, at least about 20%; even more typically, at least about 25%; even more typically, at least about 50%. Sequence identity alone as a measure of similarity is most useful when the query sequence is usually, at least 80 residues in length; more usually, 90 residues; even more usually, at least 95 amino acid residues in length. More typically, similarity can be concluded based on sequence identity alone when the query sequence is preferably 100 residues in length; more preferably, 120 residues in length; even more preferably, 150 amino acid residues in length.

It is apparent, when studying protein sequence families, that some regions have been better conserved than others during evolution. These regions are generally important for the function of a protein and/or for the maintenance of its three- dimensional structure. By analyzing the constant and variable properties of such groups of similar sequences, it is possible to derive a signature for a protein family or domain, which distinguishes its members from all other unrelated proteins. A pertinent analogy is the use of fingerprints by the police for identification purposes. A fingerprint is generally sufficient to identify a given individual. Similarly, a protein

signature can be used to assign a new sequence to a specific family of proteins and thus to formulate hypotheses about its function. The PROSITE database is a compendium of such fingerprints (motifs) and may be used with search software such as Wisconsin GCG Motifs to find motifs or fingerprints in query sequences. PROSITE currently contains signatures specific for about a thousand protein families or domains. Each of these signatures comes with documentation providing background information on the structure and function of these proteins (Hofmann *et al.* (1999) Nucleic Acids Res. **27**:215-219; Bucher and Bairoch ., A generalized profile syntax for biomolecular sequences motifs and its function in automatic sequence interpretation (In) ISMB-94; Proceedings 2nd International Conference on Intelligent Systems for Molecular Biology; Altman *et al.* Eds. (1994), pp 53-61, AAAI Press, Menlo Park).

Translations of the provided nucleic acids can be aligned with amino acid profiles that define either protein families or common motifs. Also, translations of the provided nucleic acids can be aligned to multiple sequence alignments (MSA) comprising the polypeptide sequences of members of protein families or motifs. Similarity or identity with profile sequences or MSAs can be used to determine the activity of the gene products (e.g., polypeptides) encoded by the provided nucleic acids or corresponding cDNA or genes.

Profiles can designed manually by (1) creating an MSA, which is an alignment of the amino acid sequence of members that belong to the family and (2) constructing a statistical representation of the alignment. Such methods are described, for example, in Birney et al., Nucl. Acid Res. (1996) 24(14): 2730-2739. MSAs of some protein families and motifs are available for downloading to a local server. For example, the PFAM database with MSAs of 547 different families and motifs, and the software (HMMER) to search the PFAM database may be downloaded from ftp://ftp.genetics.wustl.edu/pub/eddy/pfam-4.4/ to allow secure searches on a local server. Pfam is a database of multiple alignments of protein domains or conserved protein regions., which represent evolutionary conserved structure that has implications for the protein's function (Sonnhammer *et al.* (1998) Nucl. Acid Res. **26**:320-322; Bateman *et al.* (1999) Nucleic Acids Res. **27**:260-262).

The 3D_ali databank (Pasarella, S. and Argos, P. (1992) Prot. Engineering 5:121-137) was constructed to incorporate new protein structural and sequence data.

The databank has proved useful in many research fields such as protein sequence and structure analysis and comparison, protein folding, engineering and design and evolution. The collection enhances present protein structural knowledge by merging information from proteins of similar main-chain fold with homologous primary structures taken from large databases of all known sequences. 3D_ali databank files may be downloaded to a secure local server from http://www.embl-heidelberg.de/argos/ali/ali_form.html.

The identify and function of the gene that correlates to a nucleic acid described herein can be determined by screening the nucleic acids or their corresponding amino acid sequences against profiles of protein families. Such profiles focus on common structural motifs among proteins of each family. Publicly available profiles are known in the art.

In comparing a novel nucleic acid with known sequences, several alignment tools are available. Examples include PileUp, which creates a multiple sequence alignment, and is described in Feng et al., J. Mol. Evol. (1987) 25:351. Another method, GAP, uses the alignment method of Needleman et al., J. Mol. Biol. (1970) 48:443. GAP is best suited for global alignment of sequences. A third method, BestFit, functions by inserting gaps to maximize the number of matches using the local homology algorithm of Smith *et al.* (1981) Adv. Appl. Math. **2**:482.

IDENTIFICATION OF SECRETED & MEMBRANE-BOUND POLYPEPTIDES

Secreted and membrane-bound polypeptides of the present invention are of interest. Because both secreted and membrane-bound polypeptides comprise a fragment of contiguous hydrophobic amino acids, hydrophobicity predicting algorithms can be used to identify such polypeptides. A signal sequence is usually encoded by both secreted and membrane-bound polypeptide genes to direct a polypeptide to the surface of the cell. The signal sequence usually comprises a stretch of hydrophobic residues. Such signal sequences can fold into helical structures. Membrane-bound polypeptides typically comprise at least one transmembrane region that possesses a stretch of hydrophobic amino acids that can transverse the membrane. Some transmembrane regions also exhibit a helical structure. Hydrophobic fragments within a polypeptide can be identified by using computer algorithms. Such algorithms include Hopp & Woods, Proc. Natl. Acad. Sci.

USA (1981) 78:3824-3828; Kyte & Doolittle, J. Mol. Biol. (1982) 157: 105-132; and RAOAR algorithm, Degli Esposti et al., Eur. J. Biochem. (1990) 190: 207-219.

Another method of identifying secreted and membrane-bound polypeptides is to translate the nucleic acids of the invention in all six frames and determine if at least 8 contiguous hydrophobic amino acids are present. Those translated polypeptides with at least 8; more typically, 10; even more typically, 12 contiguous hydrophobic amino acids are considered to be either a putative secreted or membrane bound polypeptide. Hydrophobic amino acids include alanine, glycine, histidine, isoleucine, leucine, lysine, methionine, phenylalanine, proline, threonine, tryptophan, tyrosine, and valine.

IDENTIFICATION OF THE FUNCTION OF AN EXPRESSION PRODUCT

The biological function of the encoded gene product of the invention may be determined by empirical or deductive methods. One promising avenue, termed phylogenomics, exploits the use of evolutionary information to facilitate assignment of gene function. The approach is based on the idea that functional predictions can be greatly improved by focusing on how genes became similar in sequence during evolution instead of focusing on the sequence similarity itself. One of the major efficiencies that has emerged from plant genome research to date is that a large percentage of higher plant genes can be assigned some degree of function by comparing them with the sequences of genes of known function.

Alternatively, "reverse genetics" is used to identify gene function. Large collections of insertion mutants are available for *Arabidopsis*, maize, petunia, and snapdragon. These collections can be screened for an insertional inactivation of any gene by using the polymerase chain reaction (PCR) primed with oligonucleotides based on the sequences of the target gene and the insertional mutagen. The presence of an insertion in the target gene is indicated by the presence of a PCR product. By multiplexing DNA samples, hundreds of thousands of lines can be screened and the corresponding mutant plants can be identified with relatively small effort. Analysis of the phenotype and other properties of the corresponding mutant will provide an insight into the function of the gene.

In one method of the invention, the gene function in a transgenic Arabidopsis plant is assessed with anti-sense constructs. A high degree of gene duplication is

apparent in Arabidopsis, andmany of the gene duplications in Arabidopsis are very tightly linked. Large numbers of transgenic Arabidopsis plants can be generated by infecting flowers with Agrobacterium tumefaciens containing an insertional mutagen, a method of gene silencing based on producing double-stranded RNA from bidirectional transcription of genes in transgenic plants can be broadly useful for high-throughput gene inactivation (Clough and Bent (1999) Plant J. **17**; Waterhouse *et al.* (1998) Proc. Natl. Acad. Sci. U.S.A. **95**:13959). This method may use promoters that are expressed in only a few cell types or at a particular developmental stage or in response to an external stimulus. This could significantly obviate problems associated with the lethality of some mutations.

Virus-induced gene silencing may also find use for suppressing gene function. This method exploits the fact that some or all plants have a surveillance system that can specifically recognize viral nucleic acids and mount a sequence-specific suppression of viral RNA accumulation. By inoculating plants with a recombinant virus containing part of a plant gene, it is possible to rapidly silence the endogenous plant gene.

Antisense nucleic acids are designed to specifically bind to RNA, resulting in the formation of RNA-DNA or RNA-RNA hybrids, with an arrest of DNA replication, reverse transcription or messenger RNA translation. Antisense nucleic acids based on a selected nucleic acid sequence can interfere with expression of the corresponding gene. Antisense nucleic acids are typically generated within the cell by expression from antisense constructs that contain the antisense strand as the transcribed strand. Antisense nucleic acids based on the disclosed nucleic acids will bind and/or interfere with the translation of mRNA comprising a sequence complementary to the antisense nucleic acid. The expression products of control cells and cells treated with the antisense construct are compared to detect the protein product of the gene corresponding to the nucleic acid upon which the antisense construct is based. The protein is isolated and identified using routine biochemical methods.

As an alternative method for identifying function of the gene corresponding to a nucleic acid disclosed herein, dominant negative mutations are readily generated for corresponding proteins that are active as homomultimers. A mutant polypeptide will interact with wild-type polypeptides (made from the other allele) and form a non-

functional multimer. Thus, a mutation is in a substrate-binding domain, a catalytic domain, or a cellular localization domain. Preferably, the mutant polypeptide will be overproduced. Point mutations are made that have such an effect. In addition, fusion of different polypeptides of various lengths to the terminus of a protein can yield dominant negative mutants. General strategies are available for making dominant negative mutants (see for example, Herskowitz (1987) Nature 329:219). Such techniques can be used to create loss of function mutations, which are useful for determining protein function.

Another approach for discovering the function of genes utilizes gene chips and microarrays. DNA sequences representing all the genes in an organism can be placed on miniature solid supports and used as hybridization substrates to quantitate the expression of all the genes represented in a complex mRNA sample. This information is used to provide extensive databases of quantitative information about the degree to which each gene responds to pathogens, pests, drought, cold, salt, photoperiod, and other environmental variation. Similarly, one obtains extensive information about which genes respond to changes in developmental processes such as germination and flowering. One can therefore determine which genes respond to the phytohormones, growth regulators, safeners, herbicides, and related agrichemicals. These databases of gene expression information provide insights into the "pathways" of genes that control complex responses. The accumulation of DNA microarray or gene chip data from many different experiments creates a powerful opportunity to assign functional information to genes of otherwise unknown function. The conceptual basis of the approach is that genes that contribute to the same biological process will exhibit similar patterns of expression. Thus, by clustering genes based on the similarity of their relative levels of expression in response to diverse stimuli or developmental or environmental conditions, it is possible to assign functions to many genes based on the known function of other genes in the cluster.

CONSTRUCTION OF POLYPEPTIDES OF THE INVENTION AND VARIANTS THEREOF

The polypeptides of the invention include those encoded by the disclosed nucleic acids. These polypeptides can also be encoded by nucleic acids that, by virtue of the degeneracy of the genetic code, are not identical in sequence to the disclosed nucleic acids. Thus, the invention includes within its scope a polypeptide

encoded by a nucleic acid having the sequence of any one of SEQ ID NOS: 1-900 or a variant thereof.

In general, the term "polypeptide" as used herein refers to both the full length polypeptide encoded by the recited nucleic acid, the polypeptide encoded by the gene represented by the recited nucleic acid, as well as portions or fragments thereof. "Polypeptides" also includes variants of the naturally occurring proteins, where such variants are homologous or substantially similar to the naturally occurring protein, and can be of an origin of the same or different species as the naturally occurring protein. In general, variant polypeptides have a sequence that has at least about 80%, usually at least about 90%, and more usually at least about 98% sequence identity with a differentially expressed polypeptide of the invention, as measured by BLAST using the parameters described above. The variant polypeptides can be naturally or non-naturally glycosylated, i.e., the polypeptide has a glycosylation pattern that differs from the glycosylation pattern found in the corresponding naturally occurring protein.

In general, the polypeptides of the subject invention are provided in a non-naturally occurring environment, *e.g.* are separated from their naturally occurring environment. In certain embodiments, the subject protein is present in a composition that is enriched for the protein as compared to a control. As such, purified polypeptide is provided, where by purified is meant that the protein is present in a composition that is substantially free of non-differentially expressed polypeptides, where by substantially free is meant that less than 90%, usually less than 60% and more usually less than 50% of the composition is made up of non-differentially expressed polypeptides.

Also within the scope of the invention are variants; variants of polypeptides include mutants, fragments, and fusions. Mutants can include amino acid substitutions, additions or deletions. The amino acid substitutions can be conservative amino acid substitutions or substitutions to eliminate non-essential amino acids, such as to alter a glycosylation site, a phosphorylation site or an acetylation site, or to minimize misfolding by substitution or deletion of one or more cysteine residues that are not necessary for function. Conservative amino acid substitutions are those that preserve the general charge, hydrophobicity/hydrophilicity, and/or steric bulk of the amino acid substituted.

Variants also include fragments of the polypeptides disclosed herein, particularly biologically active fragments and/or fragments corresponding to functional domains. Fragments of interest will typically be at least about 10 amino acids (aa) to at least about 15 aa in length, usually at least about 50 aa in length, and can be as long as 300 aa in length or longer, but will usually not exceed about 1000 aa in length, where the fragment will have a stretch of amino acids that is identical to a polypeptide encoded by a nucleic acid having a sequence of any SEQ ID NOS:1-900, or a homolog thereof.

The protein variants described herein are encoded by nucleic acids that are within the scope of the invention. The genetic code can be used to select the appropriate codons to construct the corresponding variants.

LIBRARIES AND ARRAYS

In general, a library of biopolymers is a collection of sequence information, which information is provided in either biochemical form (e.g., as a collection of nucleic acid or polypeptide molecules), or in electronic form (e.g., as a collection of genetic sequences stored in a computer-readable form, as in a computer system and/or as part of a computer program). The term biopolymer, as used herein, is intended to refer to polypeptides, nucleic acids, and derivatives thereof, which molecules are characterized by the possession of genetic sequences either corresponding to, or encoded by, the sequences set forth in the provided sequence list (seqlist). The sequence information can be used in a variety of ways, e.g., as a resource for gene discovery, as a representation of sequences expressed in a selected cell type, *e.g.* cell type markers, *etc.*

The nucleic acid libraries of the subject invention include sequence information of a plurality of nucleic acid sequences, where at least one of the nucleic acids has a sequence of any of SEQ ID NOS:1-900. By plurality is meant one or more, usually at least 2 and can include up to all of SEQ ID NOS:1-900. The length and number of nucleic acids in the library will vary with the nature of the library, e.g., if the library is an oligonucleotide array, a cDNA array, a computer database of the sequence information, etc.

Where the library is an electronic library, the nucleic acid sequence information can be present in a variety of media. "Media" refers to a manufacture,

other than an isolated nucleic acid molecule, that contains the sequence information of the present invention. Such a manufacture provides the sequences or a subset thereof in a form that can be examined by means not directly applicable to the sequence as it exists in a nucleic acid. For example, the nucleotide sequence of the present invention, e.g. the nucleic acid sequences of any of the nucleic acids of SEQ ID NOS:1-900, can be recorded on computer readable media, *e.g.* any medium that can be read and accessed directly by a computer. Such media include, but are not limited to: magnetic storage media, such as a floppy disc, a hard disc storage medium, and a magnetic tape; optical storage media such as CD-ROM; electrical storage media such as RAM and ROM; and hybrids of these categories such as magnetic/optical storage media. One of skill in the art can readily appreciate how any of the presently known computer readable mediums can be used to create a manufacture comprising a recording of the present sequence information. "Recorded" refers to a process for storing information on computer readable medium, using any such methods as known in the art. Any convenient data storage structure can be chosen, based on the means used to access the stored information. A variety of data processor programs and formats can be used for storage, *e.g.* word processing text file, database format, etc. In addition to the sequence information, electronic versions of the libraries of the invention can be provided in conjunction or connection with other computer-readable information and/or other types of computer-readable files (e.g., searchable files, executable files, etc, including, but not limited to, for example, search program software, etc.)

By providing the nucleotide sequence in computer readable form, the information can be accessed for a variety of purposes. Computer software to access sequence information is publicly available. For example, the BLAST (Altschul et al., supra.) and BLAZE (Brutlag et al. Comp. Chem. (1993) 17:203) search algorithms on a Sybase system can be used identify open reading frames (ORFs) within the genome that contain homology to ORFs from other organisms.

As used herein, "a computer-based system" refers to the hardware means, software means, and data storage means used to analyze the nucleotide sequence information of the present invention. The minimum hardware of the computer-based systems of the present invention comprises a central processing unit (CPU), input means, output means, and data storage means. A skilled artisan can readily

appreciate that any one of the currently available computer-based system are suitable for use in the present invention. The data storage means can comprise any manufacture comprising a recording of the present sequence information as described above, or a memory access means that can access such a manufacture.

"Search means" refers to one or more programs implemented on the computer-based system, to compare a target sequence or target structural motif with the stored sequence information. Search means are used to identify fragments or regions of the genome that match a particular target sequence or target motif. A variety of known algorithms are publicly known and commercially available, e.g. MacPattern (EMBL), BLASTN, BLASTX (NCBI) and tBLASTX. A "target sequence" can be any DNA or amino acid sequence of six or more nucleotides or two or more amino acids, preferably from about 10 to 100 amino acids or from about 30 to 300 nucleotide residues.

A "target structural motif," or "target motif," refers to any rationally selected sequence or combination of sequences in which the sequence(s) are chosen based on a three-dimensional configuration that is formed upon the folding of the target motif, or on consensus sequences of regulatory or active sites. There are a variety of target motifs known in the art. Protein target motifs include, but arc not limited to, enzyme active sites and signal sequences. Nucleic acid target motifs include, but are not limited to, hairpin structures, promoter sequences and other expression elements such as binding sites for transcription factors.

A variety of structural formats for the input and output means can be used to input and output the information in the computer-based systems of the present invention. One format for an output means ranks fragments of the genome possessing varying degrees of homology to a target sequence or target motif. Such presentation provides a skilled artisan with a ranking of sequences and identifies the degree of sequence similarity contained in the identified fragment.

A variety of comparing means can be used to compare a target sequence or target motif with the data storage means to identify sequence fragments of the genome. A skilled artisan can readily recognize that any one of the publicly available homology search programs can be used as the search means for the computer based systems of the present invention.

As discussed above, the "library" of the invention also encompasses biochemical libraries of the nucleic acids of SEQ ID NOS:1-900, e.g., collections of nucleic acids representing the provided nucleic acids. The biochemical libraries can take a variety of forms, e.g. a solution of cDNAs, a pattern of probe nucleic acids stably bound to a surface of a solid support (microarray) and the like. By array is meant an article of manufacture that has a solid support or substrate with one or more nucleic acid targets on one of its surfaces, where the number of distinct nucleic may be in the hundreds, thousand, or tens of thousands. Each nucleic acid will comprise at 18 nt and often at least 25 nt, and often at least 100 to 1000 nucleotides, and may represent up to a complete coding sequence or cDNA.. A variety of different array formats have been developed and are known to those of skill in the art. The arrays of the subject invention find use in a variety of applications, including gene expression analysis, drug screening, mutation analysis and the like, as disclosed in the above-listed exemplary patent documents.

In addition to the above nucleic acid libraries, analogous libraries of polypeptides are also provided, where the where the polypeptides of the library will represent at least a portion of the polypeptides encoded by SEQ ID NOS:1-900.

GENETICALLY ALTERED CELLS AND TRANSGENICS

The subject nucleic acids can be used to create genetically modified and transgenic organisms, usually plant cells and plants, which may be monocots or dicots. The term transgenic, as used herein, is defined as an organism into which an exogenous nucleic acid construct has been introduced, generally the exogenous sequences are stably maintained in the genome of the organism. Of particular interest are transgenic organisms where the genomic sequence of germ line cells has been stably altered by introduction of an exogenous construct.

Typically, the transgenic organism is altered in the genetic expression of the introduced nucleotide sequences as compared to the wild-type, or unaltered organism. For example, constructs that provide for over-expression of a targeted sequence, sometimes referred to as a "knock-in", provide for increased levels of the gene product. Alternatively, expression of the targeted sequence can be down-regulated or substantially eliminated by introduction of a "knock-out" construct, which

may direct transcription of an anti-sense RNA that blocks expression of the naturally occurring mRNA, by deletion of the genomic copy of the targeted sequence, *etc.*

In one method, large numbers of genes are simultaneously introduced in order to explore the genetic basis of complex traits, for example by making plant artificial chromosome (PLAC) libraries. The centromeres in *Arabidopsis* have been mapped and current genome sequencing efforts will extend through these regions. Because *Arabidopsis* telomeres are very similar to those in yeast one may use a hybrid sequence of alternating plant and yeast sequences that function in both types of organisms, developing yeast artificial chromosome-PLAC libraries, and then introducing them into a suitable plant host to evaluate the phenotypic consequences. By providing a defined chromosomal environment for cloned genes, the use of PLACs may also enhance the ability to produce transgenic plants with defined levels of gene expression.

It has been found in many organisms that there is significant redundancy in the representation of genes in a genome. That is, a particular gene function is likely by represented by multiple copies of similar coding sequences in the genome. These copies are typically conserved in the amino acid sequence, but may diverge in the sequence of non-translated sequences, and in their codon usage. In order to knock out a particular genetic function in an organism, it may not be sufficient to delete a genomic copy of a single gene. In such cases it may be preferable to achieve a genetic knock-out with an anti-sense construct, particularly where the sequence is aligned with the coding portion of the mRNA.

Methods of transforming plant cells are well-known in the art, and include protoplast transformation, tungsten whiskers (Coffee et al., U.S. Pat. No. 5,302,523, issued Apr. 12, 1994), directly by microorganisms with infectious plasmids, use of transposons (U.S. Patent No. 5,792,294), infectious viruses, the use of liposomes, microinjection by mechanical or laser beam methods, by whole chromosomes or chromosome fragments, electroporation, silicon carbide fibers, and microprojectile bombardment.

For example, one may utilize the biolistic bombardment of meristem tissue, at a very early stage of development, and the selective enhancement of transgenic sectors toward genetic homogeneity, in cell layers that contribute to germline transmission. Biolistics-mediated production of fertile, transgenic maize is described

in Gordon-Kamm *et al.* (1990), Plant Cell **2**:603; Fromm *et al.* (1990) Bio/Technology **8**: 833, for example. Alternatively, one may use a microorganism, including but not limited to, *Agrobacterium tumefaciens* as a vector for transforming the cells, particularly where the targeted plant is a dicotyledonous species. See, for example, U.S. Patent No. 5,635,381. Leung *et al.* (1990) Curr. Genet. **17**(5):409-11 describe integrative transformation of three fertile hermaphroditic strains of *Arabidopsis thaliana* using plasmids and cosmids that contain an *E. coli* gene linked to *Aspergillus nidulans* regulatory sequences.

Preferred expression cassettes for cereals may include promoters that are known to express exogenous DNAs in corn cells. For example, the Adhl promoter has been shown to be strongly expressed in callus tissue, root tips, and developing kernels in corn. Promoters that are used to express genes in corn include, but are not limited to, a plant promoter such as the, CaMV 35S promoter (Odell et al., Nature, 313, 810 (1985)), or others such as CaMV 19S (Lawton et al., Plant Mol. Biol., 9, 31F (1987)), nos (Ebert et al., PNAS USA, 84, 5745 (1987)), Adh (Walker et al., PNAS USA, 84, 6624 (1987)), sucrose synthase (Yang et al., PNAS USA, 87, 4144 (1990)), .alpha.-tubulin, ubiquitin, actin (Wang et al., Mol. Cell. Biol., 12, 3399 (1992)), cab (Sullivan et al., Mol. Gen. Genet, 215, 431 (1989)), PEPCase (Hudspeth et al., Plant Mol. Biol., 12, 579 (1989)), or those associated with the R gene complex (Chandler et al., The Plant Cell, 1, 1175 (1989)). Other promoters useful in the practice of the invention are known to those of skill in the art.

Tissue-specific promoters, including but not limited to, root-cell promoters (Conkling et al., Plant Physiol., 93, 1203 (1990)), and tissue-specific enhancers (Fromm et al., The Plant Cell, 1, 977 (1989)) are also contemplated to be particularly useful, as are inducible promoters such as water-stress-, ABA- and turgor-inducible promoters (Guerrero et al., Plant Molecular Biology, 15, 11-26)), and the like.

Regulating and/or limiting the expression in specific tissues may be functionally accomplished by introducing a constitutively expressed gene (all tissues) in combination with an antisense gene that is expressed only in those tissues where the gene product is not desired. Expression of an antisense transcript of this preselected DNA segment in an rice grain, using, for example, a zein promoter, would prevent accumulation of the gene product in seed. Hence the protein encoded by the preselected DNA would be present in all tissues except the kernel.

Alternatively, one may wish to obtain novel tissue-specific promoter sequences for use in accordance with the present invention. To achieve this, one may first isolate cDNA clones from the tissue concerned and identify those clones which are expressed specifically in that tissue, for example, using Northern blotting or DNA microarrays. Ideally, one would like to identify a gene that is not present in a high copy number, but which gene product is relatively abundant in specific tissues. The promoter and control elements of corresponding genomic clones may then be localized using the techniques of molecular biology known to those of skill in the art. Alternatively, promoter elements can be identified using enhancer traps based on T-DNA and/or transposon vector systems (see, for example, Campisi *et al.* (1999) Plant J. **17**:699-707; Gu *et al.* (1998) Development **125**:1509-1517).

In some embodiments of the present invention expression of a DNA segment in a transgenic plant will occur only in a certain time period during the development of the plant. Developmental timing is frequently correlated with tissue specific gene expression. For example, in corn expression of zein storage proteins is initiated in the endosperm about 15 days after pollination.

Ultimately, the most desirable DNA segments for introduction into a plant genome may be homologous genes or gene families which encode a desired trait (e.g., increased disease resistance) and which are introduced under the control of novel promoters or enhancers, etc., or perhaps even homologous or tissue-specific (e.g., root-, grain- or leaf-specific) promoters or control elements.

The genetically modified cells are screened for the presence of the introduced genetic material. The cells may be used in functional studies, drug screening, *etc.*, *e.g.* to study chemical mode of action, to determine the effect of a candidate agent on pathogen growth, infection of plant cells, *etc.*

The modified cells are useful in the study of genetic function and regulation, for alteration of the cellular metabolism, and for screening compounds that may affect the biological function of the gene or gene product. For example, a series of small deletions and/or substitutions may be made in the host's native gene to determine the role of different domains and motifs in the biological function. Specific constructs of interest include anti-sense, as previously described, which will reduce or abolish expression, expression of dominant negative mutations, and over-expression of genes.

Where a sequence is introduced, the introduced sequence may be either a complete or partial sequence of a gene native to the host, or may be a complete or partial sequence that is exogenous to the host organism, *e.g.*, an *A. thaliana* sequence inserted into wheat plants. A detectable marker, such as aldA, *lac Z*, *etc.* may be introduced into the locus of interest, where upregulation of expression will result in an easily detected change in phenotype.

One may also provide for expression of the gene or variants thereof in cells or tissues where it is not normally expressed, at levels not normally present in such cells or tissues, or at abnormal times of development, during sporulation, *etc.* By providing expression of the protein in cells in which it is not normally produced, one can induce changes in cell behavior.

DNA constructs for homologous recombination will comprise at least a portion of the provided gene or of a gene native to the species of the host organism, wherein the gene has the desired genetic modification(s), and includes regions of homology to the target locus (see Kempin *et al.* (1997) Nature **389**:802-803). DNA constructs for random integration or episomal maintenance need not include regions of homology to mediate recombination. Conveniently, markers for positive and negative selection are included. Methods for generating cells having targeted gene modifications through homologous recombination are known in the art.

Embodiments of the invention provide processes for enhancing or inhibiting synthesis of a protein in a plant by introducing a provided nucleic acids sequence into a plant cell, where the nucleic acid comprises sequences encoding a protein of interest. For example, enhanced resistance to pathogens may be achieved by inserting a nucleic acid encoding an activator in a vector downstream from a promoter sequence capable of driving constitutive high-level expression in a plant cell. When grown into plants, the transgenic plants exhibit increased synthesis of resistance proteins, and increased resistance to pathogens.

Other embodiments of the invention provide processes for enhancing or inhibiting synthesis of a tolerance factor in a plant by introducing a nucleic acid of the invention into a plant cell, where the nucleic acid comprises sequences encoding a tolerance factor. For example, enhanced tolerance to an environmental stress may be achieved by inserting a nucleic acid encoding an activator in a vector downstream from a promoter sequence capable of driving constitutive high-level expression in a

plant cell. When grown into plants, the transgenic plants exhibit increased synthesis of tolerance proteins, and increased tolerance to environmental stress.

Factors which are involved, directly or indirectly in biosynthetic pathways whose products are of commercial, nutritional, or medicinal value include any factor, usually a protein or peptide, which regulates such a biosynthetic pathway (e.g., an activator or repressor); which is an intermediate in such a biosynthetic pathway; or which is a product that increases the nutritional value of a food product; a medicinal product; or any product of commercial value and/or research interest. Plant and other cells may be genetically modified to enhance a trait of interest, by upregulating or down-regulating factors in a biosynthetic pathway.

SCREENING ASSAYS

The polypeptides encoded by the provided nucleic acid sequences, and cells genetically altered to express such sequences, are useful in a variety of screening assays to determine effect of candidate inhibitors, activators., or modifiers of the gene product. One may determine what insecticides, fungicides and the like have an enhancing or synergistic activity with a gene. Alternatively, one may screen for compounds that mimic the activity of the protein. Similarly, the effect of activating agents may be used to screen for compounds that mimic or enhance the activation of proteins. Candidate inhibitors of a particular gene product are screened by detecting decreased from the targeted gene product.

The screening assays may use purified target macromolecules to screen large compound libraries for inhibitory drugs; or the purified target molecule may be used for a rational drug design program, which requires first determining the structure of the macromolecular target or the structure of the macromolecular target in association with its customary substrate or ligand. This information is then used to design compounds which must be synthesized and tested further. Test results are used to refine the molecular models and drug design process in an iterative fashion until a lead compound emerges.

Drug screening may be performed using an *in vitro* model, a genetically altered cell, or purified protein. One can identify ligands or substrates that bind to, modulate or mimic the action of the target genetic sequence or its product. A wide variety of assays may be used for this purpose, including labeled *in vitro* protein-

protein binding assays, electrophoretic mobility shift assays, immunoassays for protein binding, and the like. The purified protein may also be used for determination of three-dimensional crystal structure, which can be used for modeling intermolecular interactions.

Where the nucleic acid encodes a factor involved in a biosynthetic pathway, as described above, it may be desirable to identify factors, *e.g.*, protein factors, which interact with such factors. One can identify interacting factors, ligands, substrates that bind to, modulate or mimic the action of the target genetic sequence or its product. A wide variety of assays may be used for this purpose, including labeled *in vitro* protein-protein binding assays, electrophoretic mobility shift assays, immunoassays for protein binding, and the like. *In vivo* assays for protein-protein interactions in *E. coli* and yeast cells are also well-established (see Hu *et al.* (2000) Methods **20**:80-94; and Bai and Elledge (1997) Methods Enzymol. **283**:141-156).

The purified protein may also be used for determination of three-dimensional crystal structure, which can be used for modeling intermolecular interactions. It may also be of interest to identify agents that modulate the interaction of a factor identified as described above with a factor encoded by a nucleic acid of the invention. Drug screening can be performed to identify such agents. For example, a labeled in vitro protein-protein binding assay can be used, which is conducted in the presence and absence of an agent being tested.

The term "agent" as used herein describes any molecule, *e.g.* protein or pharmaceutical, with the capability of altering or mimicking a physiological function. Generally a plurality of assay mixtures are run in parallel with different agent concentrations to obtain a differential response to the various concentrations. Typically, one of these concentrations serves as a negative control, *i.e.* at zero concentration or below the level of detection.

Candidate agents encompass numerous chemical classes, though typically they are organic molecules, preferably small organic compounds having a molecular weight of more than 50 and less than about 2,500 daltons. Candidate agents comprise functional groups necessary for structural interaction with proteins, particularly hydrogen bonding, and typically include at least an amine, carbonyl, hydroxyl or carboxyl group, preferably at least two of the functional chemical groups. The candidate agents often comprise cyclical carbon or heterocyclic structures

and/or aromatic or polyaromatic structures substituted with one or more of the above functional groups. Candidate agents are also found among biomolecules including peptides, saccharides, fatty acids, steroids, purines, pyrimidines, derivatives, structural analogs or combinations thereof.

Candidate agents are obtained from a wide variety of sources including libraries of synthetic or natural compounds. For example, numerous means are available for random and directed synthesis of a wide variety of organic compounds and biomolecules, including expression of randomized oligonucleotides and oligopeptides. Alternatively, libraries of natural compounds in the form of bacterial, fungal, plant and organism extracts are available or readily produced. Additionally, natural or synthetically produced libraries and compounds are readily modified through conventional chemical, physical and biochemical means, and may be used to produce combinatorial libraries. Known pharmacological agents may be subjected to directed or random chemical modifications, such as acylation, alkylation, esterification, amidification, *etc.* to produce structural analogs.

Where the screening assay is a binding assay, one or more of the molecules may be joined to a label, where the label can directly or indirectly provide a detectable signal. Various labels include radioisotopes, fluorescers, chemiluminescers, enzymes, specific binding molecules, particles, *e.g.* magnetic particles, and the like. Specific binding molecules include pairs, such as biotin and streptavidin, digoxin and antidigoxin *etc.* For the specific binding members, the complementary member would normally be labeled with a molecule that provides for detection, in accordance with known procedures.

A variety of other reagents may be included in the screening assay. These include reagents like salts, neutral proteins, e.g. albumin, detergents, etc that are used to facilitate optimal protein-protein binding and/or reduce non-specific or background interactions. Reagents that improve the efficiency of the assay, such as protease inhibitors, nuclease inhibitors, anti-microbial agents, *etc.* may be used. The mixture of components are added in any order that provides for the requisite binding. Incubations are performed at any suitable temperature, typically between 4 and 40° C. Incubation periods are selected for optimum activity, but may also be optimized to facilitate rapid high-throughput screening. Typically between 0.1 and 1 hours will be sufficient.

The compounds having the desired biological activity may be administered in an acceptable carrier to a host. The active agents may be administered in a variety of ways. Depending upon the manner of introduction, the compounds may be formulated in a variety of ways. The concentration of therapeutically active compound in the formulation may vary from about 0.01-100 wt.%.

It must be noted that as used herein and in the appended claims, the singular forms "a", "and", and "the" include plural referents unless the context clearly dictates otherwise. Thus, for example, reference to "a complex" includes a plurality of such complexes and reference to "the formulation" includes reference to one or more formulations and equivalents thereof known to those skilled in the art, and so forth.

Unless defined otherwise, all technical and scientific terms used herein have the same meaning as commonly understood to one of ordinary skill in the art to which this invention belongs. Although any methods, devices and materials similar or equivalent to those described herein can be used in the practice or testing of the invention, the preferred methods, devices and materials are now described.

All publications mentioned herein are incorporated herein by reference for the purpose of describing and disclosing, for example, the methods and methodologies that are described in the publications which might be used in connection with the presently described invention. The publications discussed above and throughout the text are provided solely for their disclosure prior to the filing date of the present application. Nothing herein is to be construed as an admission that the inventors are not entitled to antedate such disclosure by virtue of prior invention.

The following examples are put forth so as to provide those of ordinary skill in the art with a complete disclosure and description of how to make and use the subject invention, and are not intended to limit the scope of what is regarded as the invention. Efforts have been made to ensure accuracy with respect to the numbers used (*e.g.* amounts, temperature, concentrations, *etc.*) but some experimental errors and deviations should be allowed for. Unless otherwise indicated, parts are parts by weight, molecular weight is average molecular weight, temperature is in degrees Celsius, and pressure is at or near atmospheric.

EXPERIMENTAL

Cloning and Characterization of *Arabidopsis thaliana* Genes.

Following DNA isolation, sequencing was performed using the Dye Primer Sequencing protocol, below. The sequencing reactions were loaded by hand onto a 48 lane ABI 377 and run on a 36 cm gel with the 36E-2400 run module and extraction. Gel analysis was performed with ABI software.

The Phred program was used to read the sequence trace from the ABI sequencer, call the bases and produce a sequence read and a quality score for each base call in the sequence., (Ewing *et al.* (1998) Genome Research **8**:175-185; Ewing and Green (1998) Genome Research **8**:186-194.) PolyPhred may be used to detect single nucleotide polymorphisms in sequences (Kwok *et al.* (1994) Genomics **25**:615-622; Nickerson *et al.* (1997) Nucleic Acids Research **25**(14):2745-2751.)

MicroWave Plasmid Protocol: Fill Beckman 96 deep-well growth blocks with 1 ml of TB containing 50 μg of ampicillin per ml. Inoculate each well with a colony picked with a toothpick or a 96-pin tool from a glycerol stock plate. Cover the blocks with a plastic lid and tape at two ends to hold lid in place. Incubate overnight (16-24 hours depending on the host stain) at 37° C with shaking at 275 rpm in a New Brunswick platform shaker. Pellet cells by centrifugation for 20 minutes at 3250 rpm in a Beckman GS-R6K, decant TB and freeze pelleted cell in the 96 well block. Thaw blocks on the bench when ready to continue.

Prepare the MW-Tween20 solution

For four blocks:	For 16 blocks:
50ml STET/TWEEN20	200ml STET/TWEEN
2 tubes RNAse (10mg/ml,600ulea)	8 tubes RNAse
1 tube lysozyme (25mg)	4 tubes lysozyme

Pipette RNAse and Lysozyme into the corner of a beaker. Add Tween 20 solution and swirl to mix completely. Use the Multidrop (or Biohit) to add 25ul of sterile H_20 (from the L size autoclaved bottles) to each well. Resuspend the pellets by vortexing on setting 10 of the platform vortexer. Check pellets after 4 min. and repeat as necessary to resuspend completely. Use the multidrop to add 70 μl of the

freshly prepared MW-Tween 20 solution to each well. Vortex at setting 6 on the platform vortex for 15 seconds. Do not cause frothing.

Incubate the blocks at room temperature for 5 min. Place two blocks at a time in the microwave (1000 Watts) with the tape (placed on the H1 to H12 side of the block) facing away from each other and turn on at full power for 30 seconds. Rotate the blocks so that the tapes face towards each other and turn on at full power again for 30 seconds.

Immediately remove the blocks from the microwave and add 300 μl of sterile ice cold H_2O with the Multidrop. Seal the blocks with foil tape and place them in an H_20/ice bath.

Vortex the blocks on 5 for 15 seconds and leave them in the H_2O/Ice bath. Return to step 7 until all the blocks are in the ice water bath. Incubate the blocks for 15 minutes on ice. Spin the blocks for 30 minutes in the Beckman GS-6KR with GH3.8 rotor with Microplus carrier at 3250rpm.

Transfer 100 µl of the supernatant to Corning/Costar round bottom 96 well trays. Cover with foil and put into fridge if to be sequenced right away. If not to be sequenced in the next day, freeze them at –20° C.

Dye Primer Sequencing: Spin down the DP brew trays and DNA template by pulsing in the Beckman GS-6KR with GH3.8 rotor with Microplus carrier. Big Dye Primer reaction mix trays (one 96 well cycleplate (Robbins) for each nucleotide), 3 microliters of reaction mix per well.

Use twelve channel pipetter (Costar) to add 2 µl of template to one each G,A,T,C, trays for each template plate. Pulse again to get both the reaction mix and template into the bottom of the cycle plate and put them into the MJ Research DNA Tetrad (PTC-225).

Start program Dye-Primer. Dye-primer is:

96° C, 1 min 1 cycle

96° C, 10 sec.

55° C, 5 sec.

70° C, 1 min 15 cycles

96° C, 10 sec.

70° C, 1 min. 15 cycles

4° C soak

When done cycling, using the Robbins Hydra 290 add 100 µl of 100 % ethanol to the A reaction cycle plate and pool the contents of all four cycle plates into the appropriate well.

To perform ethanol precipitation: Use Hydra program 4 to add 100 µl 100% ethanol to each A tray. Use Hydra program 5 to transfer the ethanol and therefore combine the samples from plate to plate. Once the G, A, T, and C trays of each block are mixed, spin for 30 minutes at 3250 in the Beckman. Pour off the ethanol with a firm shake and blot on a paper towel before drying in the speed vac (~10 minutes or until dry). If ready to load add 3 µl dye and denature in the oven at 95° C for ~5 minutes and load 2 µl. If to store, cover with tape and store at –20°C.

Common Solutions

Terrific Broth

Per liter:

900 ml H_2O

12 g bacto tryptone

24 g bacto-yeast extract

4 ml glycerol

Shake until dissolved and then autoclave. Allow the solution to cool to 60° C or less and then add 100 ml of sterile 0.17M KH_2PO_4, 0.72M K_2HPO_4 (in the hood w/ sterile technique).

0.17M KH_2PO_4, 0.72M K_2HPO_4

Dissolve 2.31g of KH_2PO_4 and 12.54g of K_2HPO_4 in 90 ml of H_2O.

Adjust volume to 100 ml with H_2O and autoclave.

Sequence loading Dye

20 ml deionized formamide

3.6 ml dH_20

400 µl 0.5M EDTA, pH 8.0

0.2 g Blue Dextran

*Light sensitive, cover in foil or store in the dark.

STET/TWEEN

10 ml 5M NaCl

5 ml 1M Tris, pH 8.0

1 ml 0.5M EDTA., pH 8.0

25ml Tween20

Bring volume to 500 ml with H_20

The sequencing reactions are run on an ABI 377 sequencer per manufacturer's' instructions. The sequencing information obtained each run are analyzed as follows.

Sequencing reads are screened for ribosomal., mitochondrial., chloroplast or human sequence contamination.. In good sequences, vector is marked by x's. These sequences go into biolims regardless of whether or not they pass the criteria for a 'good' sequence. This criteria is >= 100 bases with phred score of >=20 and 15 of these bases adjacent to each other.

Sequencing reads that pass the criteria for good sequences are downloaded for assembly into consensus sequences (contigs). The program Phrap (copyrighted by Phil Green at University of Washington, Seattle, WA) utilizes both the Phred sequence information and the quality calls to assemble the sequencing reads. Parameters used with Phrap are determined empirically to minimize assembly of chimeric sequences and maximize differential detection of closely related members of gene families. The following parameters are used with the Phrap program to perform the assembly:

Penalty	–6	Penalty for mismatches(substitutions)
Minmatch	40	Minimum length of matching sequence to use in assembly of reads
Trim penalty	0	penalty used for identifying degenerate sequence at beginning and end of read.
Minscore	80	Minimum alignment score

Results from the Phrap analysis yield either contigs consisting of a consensus of two or more overlapping sequence reads, or singlets that are non-overlapping .

The contig and singlets assembly are further analyzed to eliminate low quality sequence utilizing a program to filter sequences based on quality scores generated by the Phred program. The threshold quality for "high quality" base calls is 20. Sequences with less than 50 contiguous high quality bases calls at the beginning of the sequence, and also at the end of the sequence are discarded. Additionally, the

maximum allowable percentage of "low quality base calls in the final sequence is 2%, otherwise the sequence is discarded.

The stand-alone BLAST programs and Genbank databases are downloaded from NCBI for use on secure servers at the Paradigm Genetics, Inc. site. The sequences from the assembly are compared to the GenBank NR database downloaded from NCBI using the gapped version (2.0) of BLASTX. BLASTX translates the DNA sequence in all six reading frames and compares it to an amino acid database. Low complexity sequences are filtered in the query sequence. (Altschul *et al.* (1997) Nucleic Acids Res **25**(17):3389-402).

Genbank sequences found in the BLASTX search with an E Value of less than $1e^{-10}$ are considered to be highly similar, and the Genbank definition lines are used to annotate the query sequences.

When no significantly similar sequences are found as a result of the BLASTX search, the query sequences are compared with the PROSITE database (Bairoch, A. (1992) PROSITE: A dictionary of sites and patterns in proteins. Nucleic Acids Research 20:2013-2018.) to locate functional motifs.

Query sequences are first translated in six reading frames using the Wisconsin GCG pepdata program (Wisconsin Package Version 10.0, Genetics Computer Group (GCG) , Madison, Wisconsin, USA.). The Wisconsin GCG motifs program (Wisconsin Package Version 10.0, Genetics Computer Group (GCG) , Madison, Wisconsin, USA.) was used to locate motifs in the peptide sequence, with no mismatches allowed. Motif names from the PROSITE results are used to annotate these query sequences.